STAGEWISE DISCRIMINATION ALGORITHMS FOR SELECTING A SUBSET OF GROUPS OF DISCRIMINANT VARIABLES

By JOHN C. EVANS[†]
N.S.W. Dept. of Agriculture
Australia

DOUGLAS S. ROBSON and STEVEN J. SCHWAGER
Cornell University
U.S.A.

BU-876-M

March 1986

SUMMARY

Three stagewise discrimination methods, one with control of Type I error, are given for use in the selection of a subset of grouped variables. Wilks's lambda is used in Rao's test at each step to decide which groups to add or drop, if any. The cases of no subsampling of sampling units and of subsampling are both considered. A multinormal example of the implementation and the performance of these methods is given. A minimal-best-subset algorithm, which selects from the best subsets of all sizes the smallest subset that retains most of the discrimination, is better than stepwise and simultaneous stepdown algorithms.

Keywords: Discriminant variables selection; Stepwise discrimination; Stepdown discrimination; Minimal-best-subset discrimination; Bayes classification; Remote sensing; Subsampling.

1. Introduction

Evans *et al.* (1985a,b,c) proposed methods for selecting a subset of variables that gives maximal accuracy in extrinsic classification of unidentified sampling units, where there are K classes and each sampling unit must be classified as being from a specified class. In some studies, it

[*Running Title*: Stagewise Selection of Discriminant Variables]

†*Present address*: P.O. Box K220, Haymarket, 2000, Australia

may be more appropriate to identify a subset of variables that maximizes the intrinsic discrimination among observations of sampling units from different classes. Also, if many variables are involved in a classification problem and the methods of Evans *et al.* (loc. cit.) become computationally prohibitive, discriminant methods of selecting a subset of variables would offer a less expensive alternative, but with the inherent risk that the subset will not be the best for classification purposes.

This paper presents stagewise methods for selecting a subset of variables that retains most of the discrimination of the full set of variables. Section 2 summarizes the usual Rao's (1973, p.556) test of additional discrimination due to adding u groups of variables to a current v groups. Sections 3, 4, and 5 give stepwise, simultaneous stepdown, and minimal-best-subset discrimination algorithms, respectively, which use Rao's test at each step to decide which groups, if any, to add or drop. These subset selection methods are given for both the case of one observation on each sampling unit and the case of subsampling, or multiple observations, on each sampling unit.

Section 6 presents an example of the implementation of the discrimination methods, compares them with all-possible-subsets discrimination and classification results, and examines how close to optimal their ultimately selected subsets are for discrimination and classification purposes. In this example, the minimal-best-subset method does better than the stepwise and stepdown methods. The example also demonstrates that none of the stagewise discrimination methods is adequate for selecting a subset of variables for classification purposes.

2. Rao's Test for Grouped Variables

2.1 The Case of One Observation per Sampling Unit

Let Y_{gh} denote variable $h \in \{1,\cdots,d\}$ of group $g \in \{1,\cdots,T\}$, $\underline{Y}_g$ =

$(Y_{g1},\cdots,Y_{gd})'$ denote the vector of d variables in group g, $\underline{Y} \equiv (\underline{Y}_1',\cdots,\underline{Y}_T')'$ denote the vector of all dT variables, and $\underline{Y}^{(v)} \equiv (\underline{Y}_{g_1}',\cdots,\underline{Y}_{g_v}')'$ denote the subvector of $\underline{Y}$ corresponding to an arbitrary subset of $v \leq T$ groups of variables $g_1,\cdots,g_v$, where $\underline{Y}_{g_J}$ ($J=1,\cdots,v$) is the vector of variables in group $g_J \in \{1,\cdots,T\}$.

As the validity of Rao's tests depends on $\underline{Y}$ being distributed multinormally, this distribution will be assumed throughout this paper, with unequal class mean vectors $\underline{\mu}_i$ ($i=1,\cdots,K$) and a common covariance matrix $\underline{\Sigma}$. The parameters $\underline{\mu}_i$ and $\underline{\Sigma}$ are estimated from observations on $N \equiv \Sigma_{i=1}^{K} r_i$ independent calibration sampling units. Let the observation vector for a sampling unit $j \in \{1,\cdots,r_i\}$ randomly selected from class $i \in \{1,\cdots,K\}$ be denoted by $\underline{y}_{ij} \equiv (\underline{y}_{1ij}',\cdots,\underline{y}_{Tij}')'$. Then $\underline{\mu}_i$ and $\underline{\Sigma}$ are estimated by the calibration mean vectors $\bar{\underline{y}}_{i\cdot}$ and the pooled mean squares and products (MSP) matrix $\underline{S} \equiv \underline{W}/(N-K)$ where

$$\underline{W} \equiv \sum_{i=1}^{K} \sum_{j=1}^{r_i} (\underline{y}_{ij} - \bar{\underline{y}}_{i\cdot})(\underline{y}_{ij} - \bar{\underline{y}}_{i\cdot})'.$$

Rao's test is based on a ratio of two Wilks's lambdas, so notation is now developed to facilitate the definition of Wilks's lambda. Let the $dv \times 1$ subvector of $\underline{y}_{ij}$ corresponding to groups g_J, $J=1,\cdots,v$, be denoted by $\underline{y}_{ij}^{(v)} \equiv (\underline{y}_{g_1 ij}',\cdots,\underline{y}_{g_v ij}')'$; the corresponding $dv \times 1$ subvectors of $\bar{\underline{y}}_{i\cdot}$ and $\underline{\mu}_i$ by $\bar{\underline{y}}_{i\cdot}^{(v)} \equiv (\bar{\underline{y}}_{g_1 i\cdot}',\cdots,\bar{\underline{y}}_{g_v i\cdot}')'$ and $\underline{\mu}_i^{(v)} \equiv (\underline{\mu}_{g_1 i}',\cdots,\underline{\mu}_{g_v i}')'$; and the corresponding $dv \times dv$ submatrices of $\underline{\Sigma}$, $\underline{S}$, and $\underline{W}$ by $\underline{\Sigma}^{(v)} \equiv \underline{\Sigma}(g_1,\cdots,g_v)$, $\underline{S}^{(v)} \equiv \underline{S}(g_1,\cdots,g_v)$, and $\underline{W}^{(v)} \equiv \underline{W}(g_1,\cdots,g_v)$. Based on $\underline{Y}^{(v)}$, Wilks's lambda is denoted by $\Lambda_v \equiv \Lambda(g_1,\cdots,g_v)$ and defined for $dv \leq N-K$ as the determinantal ratio

$$\Lambda_v \equiv |\underline{W}^{(v)}| \;/\; |\underline{W}^{(v)} + \underline{B}^{(v)}| \qquad (2.1.1)$$

where $\underline{B}^{(v)}$ is a dv × dv submatrix of $\underline{B} \equiv \Sigma_{i=1}^{K} r_i(\bar{\underline{y}}_{i.} - \bar{\underline{y}}_{..})(\bar{\underline{y}}_{i.} - \bar{\underline{y}}_{..})'$, the among classes sums of squares and products (SSP) matrix.

To perform Rao's test of additional discrimination due to adding u groups to the current v groups, proceed as follows. Calculate $\Lambda_{u\cdot v} \equiv \Lambda(g_{v+1},\cdots,g_{v+u}|g_1,\cdots,g_v)$ as

$$\Lambda_{u\cdot v} = \Lambda_{u+v} / \Lambda_v \ , \qquad (2.1.2)$$

where $\Lambda_{u+v} = \Lambda(g_1,\cdots,g_{u+v})$ is based on $\underline{Y}^{(u+v)}$, for u+v ≤ T. Then evaluate $F_{u\cdot v} \equiv F(g_{v+1},\cdots,g_{v+u}|g_1,\cdots,g_v)$ as

$$F_{u\cdot v} = (ab - 2c)\ (1 - \Lambda_{u\cdot v}^{1/b}\ )\ /\ [du(K-1)\Lambda_{u\cdot v}^{1/b}] \qquad (2.1.3)$$

where

$$a = N - K - dv - \tfrac{1}{2}(du - K + 2),$$

$$b = \begin{cases} [d^2u^2(K-1)^2 - 4]^{\frac{1}{2}} \ /\ [d^2u^2 + (K-1)^2 - 5]^{\frac{1}{2}} & \text{if } d^2u^2 + (K-1)^2 \neq 5 \\ 1 & \text{otherwise,} \end{cases}$$

and

$$c = [du(K-1) - 2]\ /\ 4.$$

By Rao (1973, p.556), $F_{u\cdot v}$ is distributed approximately as $F_{du(K-1),ab-2c}$. Rao's test is done by comparing $F_{u\cdot v}$ with $F^{\alpha}_{du(K-1),ab-2c}$, the appropriate critical value of the F distribution with du(K-1) and ab-2c d.f. for a given significance level α. If $F_{u\cdot v} > F^{\alpha}_{du(K-1),ab-2c}$, then the v groups have provided additional discrimination; otherwise they have not.

2.2 The Case of Subsampled Sampling Units

Evans *et al.* (1985c) derived methods for classifying a subsampled sampling unit on the basis of the mean vector of the subsamples (=subsampling units=observations) on this unit. Wilks's lambda and Rao's test will now be defined, analogously to Section 2.1, for the case of subsampling. It will be assumed that the number of subsamples for each

calibration sampling unit has been prespecified and that the subsamples have been selected independently. Let the observation vector on subsample $k\varepsilon\{1,\cdots,n_{ij}\}$ of calibration sampling unit $j\varepsilon\{1,\cdots,r_i\}$ from class $i\varepsilon\{1,\cdots,K\}$ be denoted by $\underline{y}_{ijk} = (\underline{y}'_{1ijk},\cdots,\underline{y}'_{Tijk})'$. As in Evans *et al.* (loc. cit.), the model assumed to be appropriate for $\underline{y}_{ijk}$ and used here is

$$\underline{y}_{ijk} = \underline{\mu}_i + \underline{\varepsilon}_{ij} + \underline{\delta}_{ijk}$$

where $\underline{\mu}_i$ is fixed but unknown, $\underline{\varepsilon}_{ij} \sim$ i.i.d. $N(\underline{0},\underline{\Sigma}_\varepsilon)$, $\underline{\delta}_{ijk} \sim$ i.i.d. $N(\underline{0},\underline{\Sigma}_\delta)$, and the $\underline{\varepsilon}_{ij}$ terms are independent of the $\underline{\delta}_{ijk}$ terms. Then it follows that $E(\underline{y}_{ijk}) = \underline{\mu}_i$, $V(\underline{y}_{ijk}) = \underline{\Sigma}_\varepsilon + \underline{\Sigma}_\delta$, $\mathrm{Cov}(\underline{y}_{ijk},\underline{y}_{ijk'}) = \underline{\Sigma}_\varepsilon$ for $k\neq k'=1,\cdots,n_{ij}$, and $\mathrm{Cov}(\underline{y}_{ijk},\underline{y}_{ij'k'}) = \underline{0}$ for $j\neq j'=1,\cdots,r_i$.

Analogously to Section 2.1, based on the observations of $N = \Sigma_{i=1}^{K} r_i$ independent calibration sampling units, define the quantities

$$\bar{\underline{y}}_{i..} = \sum_{j=1}^{r_i}\sum_{k=1}^{n_{ij}} \underline{y}_{ijk} \,/\, n_{i\cdot}$$

$$= \sum_{j=1}^{r_i} n_{ij}\, \bar{\underline{y}}_{ij\cdot}/n_{i\cdot} \quad ,$$

$$\underline{W} = \sum_{i=1}^{K}\sum_{j=1}^{r_i} n_{ij}(\bar{\underline{y}}_{ij\cdot} - \bar{\underline{y}}_{i..})(\bar{\underline{y}}_{ij\cdot} - \bar{\underline{y}}_{i..})'$$

$$= (N-K)\underline{S} \quad ,$$

and

$$\underline{B} = \sum_{i=1}^{K} n_{i\cdot}(\bar{\underline{y}}_{i..} - \bar{\underline{y}}_{...})\,(\bar{\underline{y}}_{i..} - \bar{\underline{y}}_{...})',$$

where

$$\bar{\underline{y}}_{...} = \sum_{i=1}^{K}\sum_{j=1}^{r_i}\sum_{k=1}^{n_{ij}} \underline{y}_{ijk} \,/\, n_{\cdot\cdot}$$

$$= \sum_{i=1}^{K} n_{i.} \bar{y}_{i..} / n_{..}$$

and

$$n_{i.} = \sum_{j=1}^{r_i} n_{ij} \ , \ n_{..} = \sum_{i=1}^{K} n_{i.} \quad .$$

Let the dv × 1 subvectors of y_{ijk}, $\bar{y}_{ij.}$, $\bar{y}_{i..}$, $\bar{y}_{...}$, and μ_i corresponding to an arbitrary v groups $g_1,\cdots,g_v$ be denoted by $y_{ijk}^{(v)}$, $\bar{y}_{ij.}^{(v)}$, $\bar{y}_{i..}^{(v)}$, $\bar{y}_{...}^{(v)}$, and $\mu_i^{(v)}$, respectively, and let the corresponding dv × dv submatrices of W, S, and B be denoted by $W^{(v)}$, $S^{(v)}$, and $B^{(v)}$, respectively. Then Wilks's lambda is given by Equation (2.1.1), after incorporating the new definitions of $W^{(v)}$ and $B^{(v)}$ here, and Rao's test is exactly as in Subsecton 2.1. As in that subsection, to enable the definition of a Wilks's lambda, it has been necessary to pool together the among sampling units SSP matrices from the K classes. This is in contrast to the classification algorithms of Evans *et al.* (1985 a,b,c), which have the advantage that the separate SSP matrices can be utilized if necessary in estimation and hypothesis testing.

3. **Stepwise Discrimination**

Jennrich (1977) gave a stepwise discriminant analysis that used Rao's test at each step to select a subset of single variables in the case of no subsampling. Lam and Cox (1981) extended Jennrich's algorithm from single to paired variables. Evans (1984) extended (and corrected) Lam and Cox's algorithm to the case of grouped variables with groups of arbitrary size $d \geq 1$. This general case is given here and is applicable when subsampling is either present or absent.

Step 0. Calculate $\Lambda(g_1)$ for each group $g_1=1,\cdots,T$ by Equation (2.1.1), where $\Lambda_1 = \Lambda(g_1)$ is equivalent to $\Lambda_{1\cdot 0}$ of Equation (2.1.2) with $\Lambda_0 \equiv 1$ to correspond to no discrimination with no data. Find the group g_1 with maximum $F(g_1)$ of Equation (2.1.3) and enter that group if $F(g_1) > F^{\alpha}_{d(K-1),ab-2c}$; otherwise stop.

Step 1. After selecting v groups ($1 \le v \le T-1$), calculate for each of the remaining $u = T-v$ groups g_k, $k\varepsilon\{v+1,\cdots,T\}$, the statistic $F_{1\cdot v} = F(g_k|g_1,\cdots,g_v)$ of Equation (2.1.3). Find the group g_k with maximum $F(g_k|g_1,\cdots,g_v)$ and enter it as the $(v+1)st$ group if $F(g_k|g_1,\cdots,g_v) > F^{\alpha}_{d(K-1),ab-2c}$; otherwise stop.

Step 2. Before considering the addition of a $(v+2)nd$ group ($3 \le v+2 \le T$), calculate for each of the currently entered v+1 groups $g_{k'}$, $k'\varepsilon\{1,\cdots,v+1\}$, the quantity $F_{1\cdot v} = F(g_{k'}|g_{1'},\cdots,g_{v'})$ by Equation (2.1.3), where $g_{1'},\cdots,g_{v'}$ identify the other v of the v + 1 currently entered groups. Find the group $g_{k'}$ with minimum $F(g_{k'}|g_{1'},\cdots,g_{v'})$ and remove it if its $F(g_{k'}|g_{1'},\cdots,g_{v'}) < F^{\alpha}_{d(K-1),ab-2c}$. Return to Step 1.

This procedure consists of alternations of Steps 1 and 2 and terminates when no further group can be added or dropped, or earlier if $dv > N-K$, which would cause $\underline{S}^{(v)}$ to be singular and would invalidate Equation (2.1.1).

4. Simultaneous Stepdown Discrimination

Calinski and Kaczmarek (1977) gave a simultaneous (i.e., fixed overall probability of a Type I error) stepdown algorithm that uses Rao's test at each step to assess whether to drop a variable, for the case of no subsampling. Evans (1984) presented details of Mudholkar and Subbaiah's (1980) generalization of that algorithm to the case of grouped variables, with groups of arbitrary size $d \ge 1$. This algorithm requires a prespecified order of testing of groups (in the order of increasing importance

or relevance). The main advantage of this procedure is that the individual significance level at each step is lower than the overall significance level, so less redundant groups are likely to be selected. An obvious disadvantage is that the final subset selected depends on the order of testing. The algorithm is now presented. It is applicable when subsampling is either present or absent.

Consider the factorization

$$\Lambda(1,\cdots,T) = \Lambda(g_1)\Lambda(g_2|g_1)\cdots\Lambda(g_T|g_1,\cdots,g_{T-1}) \ , \qquad (4.1)$$

which follows from repeated application of $\Lambda_{u+v} = \Lambda_v\Lambda_{u\cdot v}$ in Equation (2.1.2). Now, instead of using $\Lambda(1,\cdots,T)$ to test the null hypothesis H_0 of no discrimination among classes by the T groups, its factors can be tested separately to identify which groups, if any, lead to a rejection of H_0. That is, the factorization of $\Lambda(1,\cdots,T)$ in Equation (4.1) corresponds to a decomposition of H_0 into T component hypotheses, $H_0 = \cap_{k=1}^{T} H_{0g_k}$, where H_{0g_k} is the null hypothesis of no discrimination due to group g_k. For subsequent simplicity of notation, let $\Lambda(g_1) = \Lambda_1$ in Equation (4.1) also be defined as $\Lambda_{1\cdot 0}$ of Equation (2.1.2), with $\Lambda_0 = 1$ corresponding to no discrimination with no data, and then $F(g_1) = F_{1\cdot 0}$ of Equation (2.1.3) is the appropriate statistic for testing H_{0g_1}. Under H_0, as $\Lambda(g_k|g_1,\cdots,g_{k-1}) = \Lambda_{1\cdot k-1}$ for k=1,$\cdots$,T, with $\Lambda(g_1)$ being used when k=1, are distributed independently, so are the $F(g_k|g_1,\cdots,g_{k-1}) = F_{1\cdot k-1}$ obtained by substituting $\Lambda_{1\cdot k-1}$ in Equation (2.1.3) for k=1,$\cdots$,T, with $F(g_1)$ being the statistic when k=1. The component hypotheses are tested in the order $H_{0g_T},\cdots,H_{0g_1}$ corresponding to the order of increasing importance from g_T to g_1. For each k=T,$\cdots$,1, $F(g_k|g_1,\cdots,g_{k-1})$ is compared to

$F^{\alpha_k}_{d(K-1),ab-2c}$, where $\alpha_k = P[F(g_k|g_1,\cdots,g_{k-1}) > F^{\alpha_k}_{d(K-1),ab-2c}|H_{0g_k}]$ is a

prespecified Type I error probability. An overall Type I error probability of $\alpha \equiv 1 - \Pi_{k=1}^{T}(1-\alpha_k)$ is achieved here by setting

$$\alpha_k \equiv 1 - (1 - \alpha)^{\left(k \sum_{k'=1}^{T} \frac{1}{k'}\right)^{-1}} \quad \text{for } k=1,\cdots,T.$$

The details of stepdown discrimination are now given.

Step 1. If $F(g_T|g_1,\cdots,g_{T-1}) > F^{\alpha_T}_{d(K-1),ab-2c}$ then reject H_{0g_T} at level α_T and H_0 at level α and stop the procedure by selecting all groups. Otherwise drop group g_T and go to Step 2.

Step 2. If $F(g_{T-1}|g_1,\cdots,g_{T-2}) > F^{\alpha_{T-1}}_{d(K-1),ab-2c}$ then reject $H_{0g_{T-1}}$ at level α_{T-1} and H_0 at level α and stop the procedure by selecting groups $g_1,\cdots,g_{T-1}$. Otherwise drop group g_{T-1} and go to Step 3.

And so on.

Step T. If $F(g_1) > F^{\alpha_1}_{d(K-1),ab-2c}$ then reject H_{0g_1} at level α_1 and H_0 at level α and stop the procedure by selecting group g_1. Otherwise select no groups and stop.

If $dT > N-K$ then this algorithm can only be partially implemented by forcing the nonselection of the groups $g_T, g_{T-1}, \cdots$ for which $dT, d(T-1), \cdots > N-K$ and testing onwards from the step k at which $dk \leq N-K$ is first satisfied.

5. Minimal-Best-Subset Discrimination

The disadvantage of stepwise and stepdown discrimination is that the selected subset may not be the best of its size, or adequate compared with all groups, for discrimination purposes. A check of adequacy could be made by appending Rao's test to either algorithm. But the minimal-best-subset

algorithm to be given next performs a check of adequacy and guarantees that the selected subset is the best of its size, although it is more expensive to apply, requiring discrimination results from all possible subsets of groups of variables.

Evans *et al.* (1985a) gave the following all-possible-subsets discrimination algorithm. Calculate $\Lambda(g_1,\cdots,g_v)$ by Equation (2.1.1) for every subset of groups of size v=1,$\cdots$,T, or if dT > N-K, only up to the largest size for which $\underset{\sim}{S}^{(v)}$ is nonsingular. For each size v=1,2,$\cdots$, find the subset of groups $g_1,\cdots,g_v$ that achieves

$$\underset{g_1,\cdots,g_v \in \{1,\cdots,T\}}{\text{minimum}} \Lambda(g_1,\cdots,g_v) \quad ;$$

then each of these subsets is regarded as the best subset of its size. An algorithm is now given to find the smallest of these best subsets that retains most of the discrimination of the full set of groups. This algorithm is applicable when subsampling is either present or absent.

Evans (1984) gave the following minimal-best-subset discrimination algorithm, which involves a series of dependent Rao tests of additional discrimination and is not a simultaneous test procedure. Using Rao's test of Equation (2.1.3), the groups not included in the best subsets of size T-1,T-2,$\cdots$,1, and 0 are tested successively for their discrimination additional to that of the included groups, or until rejection of one of the null hypotheses of no additional discrimination. (The groups not included in the "best" (null) subset of size 0 are all groups, and the test of their discrimination uses $\Lambda(1,\cdots,T|0) \equiv \Lambda(1,\cdots,T)$, with $\Lambda_0 \equiv 1$ corresponding to no discrimination when there are no data.) If and when such a rejection occurs, select the groups included in the best subset at the previous step;

otherwise select no groups. This final subset is taken as the minimal-best subset. If dT > N-K then this algorithm cannot be applied, as comparisons cannot be made against the full set of groups.

An alternative approach for the comparison of all possible subsets has been advocated by McKay and Campbell (1982). That method "employs significance testing with protection of the simultaneous level of significance for all (Rao's) tests of additional information carried out" in isolating a number of adequate subsets that give essentially the same discrimination as the original set of variables. This alternative will not be implemented in the example of Section 6, as a subjective decision would have to be made, for each simulated data set there, about which one of the adequate subsets would be used to compare with the single subsets selected by the other methods.

6. A Remote Sensing Example

Background details for this example are given in Evans *et al.* (1985a). In the examples of Evans *et al.* (1985b,c), stepwise, stepdown, and minimal-best-subset classification algorithms were applied to seven simulated data sets (labelled as data sets 2-6, 8, and 9) for the case of no subsampling (Evans *et al.* 1985b) and six more for the case of subsampling (Evans *et al.* 1985c). Each of these algorithms used Evans's (1984) test of additional reduction in Bayes risk (= increase in classification accuracy) at each step to decide whether to add or drop a group of variables. The discrimination counterparts of these algorithms, described in this paper, have been applied to the same 13 data sets. Subsets selected using these discriminant analyses have been assessed for their optimality; first, for discrimination by comparing their Wilks's lambdas with those of the best

(= minimum lambda) subsets of each size; and, second, for classification by comparing their standardized estimated Bayes risks (= z-values) with those of the best (= minimum z) subsets of each size (from Evans *et al.* 1985b,c).

Each discrimination algorithm was implemented with the MATRIX procedure of the SAS package (SAS Institute Inc., 1982). All data sets were simulated from different sets of K = 5 multivariate normal distributions with a common covariance matrix and d = 4 and T = 5. From each of data sets 2, 3, and 4 of Evans *et al.* (1985b) and the 6 data sets of Evans *et al.* (1985c), the $r_i = 10 = m_i$ (N = 50, M = 50) observations are regarded here as $r_i = 20$ calibration observations from each class $i=1,\cdots,5$ (N = 100). [In Evans *et al.* (1985b,c), $M = \Sigma_{i=1}^{K} m_i$ observations were used only in the estimation of classification accuracy.] Similarly, the $r_i = 5$, $m_i = 6$ (N = 25, M = 30) observations from data sets 5, 6, 8, and 9 are used here as $r_i = 11$ (N = 55) calibration observations. The original SAS MATRIX programs for simulation and subset selection can be found in Evans (1984), and updates can be obtained from the first author of this paper. The total cost (including CPU time and other costs) of executing all discrimination analyses on an IBM 3081 under OS VS2/MVS was about $U.S. 5 (CPU time: about 4.2 seconds) per data set.

Due to the high degree of discrimination in the data and the sensitivity of Rao's test with large N, each discrimination method retained all groups for each data set using significance levels (α) of 0.05 and 0.01, thus achieving no subset selection. To force the selection of a subset of groups to enable a study of its optimality, lower significance levels of 0.001 and 0.0001 were used with the data sets (5, 6, 8, and 9) having the lower N = 55. Only the stepwise method responded by selecting no groups in some cases and all five groups in other cases, but in no case between none

and five groups. In an attempt to force the selection of between 0 and 5 groups, an N = 25 subsample of each of data sets 5, 6, 8, and 9 was analyzed using the original $\alpha = 0.05$ and 0.01 values. This was successful for the stepwise algorithm in some cases, but no groups were selected in other cases. It was also successful for the minimal-best-subset method, which selected between 0 and 5 groups in all but one case. Although the stepdown method was forced to select less than five groups as desired, it went to the other extreme by selecting no groups in most cases; using an N between 25 and 55 would probably have achieved a selection of between 0 and 5 groups in most cases. As in Evans *et al.* (1985b,c) for stepdown classification, both chronological and reverse testing order of groups (= dates) were used for the above stepdown discriminations. In all cases, the chronological stepdown method selected no groups.

[Table 1 should go about here]

For the N = 25 subsamples of data sets 5, 6, 8, and 9, Table 1 gives the groups selected using $\alpha = 0.05$ at each step of the stepwise and minimal-best-subset discriminations and an overall $\alpha = 0.05$ for the reverse chronological stepdown method. By inspection of the progressive reductions in Λ tabulated there, from the Λ-best subset of size 1 to the Λ-best of size 4, it is not obvious whether the stagewise algorithms have selected optimal subsets for discrimination. The choice of which algorithm to use for discrimination must therefore be based on other criteria. Stepdown discrimination cannot be relied upon, as it depends on testing order and does not guarantee that the selected subset is the best of its size. For two of the four data sets, the stepwise algorithm selected no groups, whereas the minimal-best-subset method retained either 4 or 5 groups. As the minimal-best-subset method has a check for adequacy of the

final subset versus all groups, it appears that the stepwise method, in selecting no groups in two cases, may in general select too few groups. Perhaps a backward stepping algorithm, which omits the worst group, if any, at each step, would retain more groups and thus be preferable (if $dT \leq N-K$) to the stepwise method used here. Based on these results, the only methods that should be considered in practice are the minimal-best-subset algorithm used here, the variation advocated by McKay and Campbell (1982), or possibly a backward stepwise method.

The results of Table 1 can also be used to assess whether any of the subsets selected by discrimination methods are anywhere near adequate for classification purposes. By inspection of the progressive reductions in estimated Bayes risk from the best subset of size 1 to the best of size 4, two main observations can be made. First, for data sets 5, 6, 8, and 9, respectively, the z-best subsets of 4, 3, 3, and 2 groups are optimal for classification. Second, not one subset selected by the discrimination algorithms was an optimal subset. By far the closest was the subset of 4 groups ($z = 3.0$) selected by the minimal-best-subset method on data set 9, for which the optimal subset consisted of only 2 groups and had a one-third lower estimated Bayes risk ($z = 2.0$). It is clear that none of the stage-wise discrimination algorithms applied here should be considered for selection of groups to use for classification purposes.

7. Conclusion

Stepwise, minimal-best-subset, and simultaneous stepdown methods have been described for the selection of discriminant variables. These procedures were applied to several multinormal data sets. Results suggested that the minimal-best-subset algorithm is better than the others. However,

more extensive future comparisons of these and alternative algorithms are necessary to establish the best procedure. As expected from previous work (Evans *et al.* 1985a,b), none of the discrimination algorithms is suitable for selecting a subset of variables for classification purposes.

References

Calinski, T. and Kaczmarek, Z. (1977) A step-down procedure of eliminating variables in multivariate analysis of variance. *Biometrical Journal*, **19**, 449-453.

Evans, J. C. (1984) Stagewise Selection and Classification of Multivariate Repeated Measurements. Ph.D. Thesis, Cornell University, 1983. Ann Arbor, Michigan: University Microfilms International.

Evans, J. C., Robson, D. S., and Schwager, S. J. (1985a) The use of estimated Bayes risk as a criterion for selecting subsets of allocation variables. Biometrics Unit Report BU-870-M. Cornell University, Ithaca, N.Y., U.S.A. Submitted for publication.

———— (1985b) Stagewise classification algorithms for selecting a subset of groups of allocation variables. Biometrics Unit Report BU-871-M. Cornell University, Ithaca, N.Y., U.S.A. Submitted for publication.

———— (1985c) Bayes classification and selection of allocation variables when sampling units are subsampled. Biometrics Unit Report BU-872-M. Cornell University, Ithaca, N.Y., U.S.A. Submitted for publication.

Jennrich, R. I. (1977) Stepwise discriminant analysis. In *Mathematical Methods for Digital Computers*, Vol. 3: *Statistical Methods for Digital Computers*, (K. Enslein, A. Ralston and H.S. Wilf, eds.), pp. 76-95. New York: Wiley.

Lam, C. F. and Cox, M. (1981) A discriminant analysis procedure for paired variables. *Technometrics*, **23**, 185-187.

McKay, R. J. and Campbell, N. A. (1982) Variable selection techniques in discriminant analysis, I. Description. *Brit. J. Math. Statist. Psych.*, **35**, 1-29.

Mudholkar, G. S. and Subbaiah, P. (1980) A review of step-down procedures for multivariate analysis. In *Multivariate Statistical Analysis*, (R.P. Gupta, ed.), pp. 161-178. Amsterdam: North Holland Publishing Company.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley.

SAS Institute Inc. (1982) *SAS User's Guide: Statistics*, 1982 Edition. Cary, North Carolina: SAS Institute Inc.

TABLE 1

The subset of groups selected and its Λ and z values for the minimal-best, stepwise, and reverse stepdown discriminations, along with Λ and z values for the Λ_v-best and z_v-best subsets of each size v=1,2,3,4.

	Data Set 5			*Data Set 6*		
	Minimal Best	*Step-wise*	*Step-down*	*Minimal Best*	*Step-wise*	*Step-down*
Groups	1,2,3,4,5	0	0	1,2,5	1,2,5	1,2,3
Λ×100				0.1	0.1	0.1
Λ-rank*				1	1	3
z				3.5	3.5	4.7
z-rank				5	5	7

	v = 1		*v = 2*		*v = 1*		*v = 2*	
	Λ-best	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*
Groups	5	1	1,5	2,5	5	5	1,5	1,5
Λ×100	32.2	37.6	4.0	6.2	23.3	23.3	1.5	1.5
Λ-rank	1	2	1	4	1	1	1	1
z	4.6	4.6	3.7	3.0	4.7	4.7	2.7	2.7
z-rank	2	1	3	1	1	1	1	1

	v = 3		*v = 4*		*v = 3*		*v = 4*	
	Λ-best	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*
Groups	1,4,5	1,3,5	1,2,3,5	1,3,4,5	1,2,5	1,3,5	1,2,3,4	1,3,4,5
Λ×100	0.3	0.3	0.0	0.0	0.1	0.1	0.0	0.0
Λ-rank	1	2	1	4	1	2	1	5
z	3.1	2.3	1.8	1.5	3.5	1.0	4.0	1.4
z-rank	4	1	3	1	5	1	5	1
	z(0) = 6.0		z(1,2,3,4,5) = 9.8		z(0) = 6.0		z(1,2,3,4,5) = 6.7	
	Λ(1,2,3,4,5) = 0.0				Λ(1,2,3,4,5) = 0.0			

TABLE 1 continued

	Data Set 8			*Data Set 9*		
	Minimal Best	*Step-wise*	*Step-down*	*Minimal Best*	*Step-wise*	*Step-down*
Groups	1	1	1	1,2,3,5	0	0
Λ×100	15.5	15.5	15.5	0.0		
Λ-rank	1	1	1	1		
z	5.1	5.1	5.1	3.0		
z-rank	2	2	2	3		

	v = 1		*v = 2*		*v = 1*		*v = 2*	
	Λ-best	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*
Groups	1	1	1,2	2,5	5	5	1,3	2,5
Λ×100	15.5	15.5	4.2	6.0	26.0	26.0	4.4	5.0
Λ-rank	1	1	1	5	1	1	1	3
z	5.1	5.1	4.6	3.1	4.8	4.8	4.2	2.0
z-rank	2	2	4	1	1	1	6	1

	v = 3		*v = 4*		*v = 3*		*v = 4*	
	Λ-best	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*	*Λ-best*	*z-best*
Groups	1,2,5	1,2,5	1,2,3,5	1,2,4,5	1,3,5	3,4,5	1,2,3,5	1,2,4,5
Λ×100	0.4	0.4	0.0	0.0	0.3	0.6	0.0	0.0
Λ-rank	1	1	1	3	1	5	1	3
z	1.9	1.9	3.0	2.7	3.1	2.6	3.0	2.4
z-rank	1	1	2	1	5	1	3	1
	z(0) = 6.0		z(1,2,3,4,5) = 4.5		z(0) = 6.0		z(1,2,3,4,5) = 2.2	
	Λ(1,2,3,4,5) = 0.0				Λ(1,2,3,4,5) = 0.0			

* Λ-rank and z-rank values are ranks from 1 (lowest=best) to 5 or 10 (highest) of Λ and z, respectively, among the 5 or 10 subsets of the same size v = 1 or 4 or v = 2 or 3. Although Λ × 100 values are given only to one decimal place in the table, rankings were done on the original values with several decimal places.